Maag Merki, Katharina; Klieme, Eckhard; Holmeier, Monika

Unterrichtsgestaltung unter den Bedingungen zentraler Abiturprüfungen. Differenzielle Analysen auf Schulebene mittels Latent Class Analysen

Zeitschrift für Pädagogik 54 (2008) 6, S. 791-808

ZEITSCHRIFT FÜR PÄDAGOGIK
Jahrgang 54 – Heft 6
November/Dezember 2008

Inhaltsverzeichnis

Thementeil: Systeme der Rechenschaftslegung und Schulentwicklung

Allgemeiner Teil

Besprechungen

Dokumentation

Katharina Maag Merki/Eckhard Klieme/Monika Holmeier

Unterrichtsgestaltung unter den Bedingungen zentraler Abiturprüfungen

Differenzielle Analysen auf Schulebene mittels Latent Class Analysen

Zusammenfassung: *Die Einführung zentraler Prüfungen ist ein wesentliches Element von Standardsicherung in den neu eingesetzten Output-Steuerungsmodellen. Inwiefern diese Prüfungen allerdings funktional sind für die Zielerreichung, ist insbesondere für die deutschsprachigen Länder empirisch nur ungenügend untersucht. In der vorliegenden Studie werden Effekte der Implementation zentraler Abiturprüfungen auf wesentliche Dimensionen der Unterrichtsqualität analysiert. Basis sind standardisierte Befragungen bei Abiturientinnen und Abiturienten in den beiden Bundesländern Hessen und Bremen im Abiturjahr 2007. Mittels Latent Class Analysen werden unterschiedliche Ergebnisstrukturen zwischen den Gymnasien und den Bundesländern herausgearbeitet, die in Teilen für das Lernen der Schüler/innen als produktiv interpretiert werden können.*

Forschungsbefunde verweisen auf die aktuell nur geringe Standardsicherung im deutschen Schulsystem und auf eine hohe diesbezügliche Varianz zwischen den Bundesländern und Schulen (vgl. Baumert/Watermann 2000; Köller u.a. 2004; Köller/Baumert/Schnabel 1999). Dies deutet darauf hin, dass zwischen den verschiedenen Schulen und Bundesländern substanzielle Unterschiede bestehen, bei gleichen Testleistungen je nach Schule bzw. Bundesland unterschiedliche Noten vergeben werden und die Kriterien für die Vergabe der Noten nicht einheitlich sind. Um Unterschiede hinsichtlich des Niveaus der fachlichen Anforderungen, der Maßstäbe zur Beurteilung von Schülerleistungen und letztendlich des erreichten Niveaus auszugleichen, werden Verfahren gefordert, die eine höhere Standardsicherung bewirken (vgl. Klieme 2004). Ein wesentliches Element von Standardsicherung im Rahmen der Output-Steuerungsmodelle ist die Einführung zentraler Prüfungen. Inwiefern dieses Steuerungskonzept allerdings funktional ist für die Zielerreichung, ist insbesondere für die deutschsprachigen Länder empirisch nur ungenügend untersucht. Die wissenschaftliche Diskussion verweist darauf, dass die Wirkung standardbezogener Zertifizierungs-, Selektions- und Allokationsprozesse hinsichtlich der Zielerreichung fraglich ist. So zeigen beispielsweise die PISA-Ergebnisse 2000 (vgl. Baumert u.a. 2003, S. 330ff.), dass es auch innerhalb der Bundesländer wie Bayern oder Baden-Württemberg, in denen nach Klasse 10 zentrale Prüfungen durchgeführt werden, zwischen den Gymnasien beachtliche Leistungsunterschiede oder Unterschiede beim Zusammenhang zwischen der Leistung und der Benotung gibt.

Zudem muss die Funktionalität in Abhängigkeit zentraler Aspekte des implementierten Monitoringkonzeptes beurteilt werden, da standardbasierte Monitoringsysteme unerwünschte Folgen wie beispielsweise teaching-to-the-test oder korrumpierende

Handlungsweisen der Lehrpersonen zur Folge haben können. Dies ist der Fall insbesondere im Kontext eines rigiden, mit starken Kontrollmechanismen, geringen Unterstützungs- und Förderleistungen sowie mit massiven negativen Konsequenzen für die Schüler/innen, Lehrpersonen oder Schulleitungen ausgestalteten Monitoringsystems (vgl. Amrein/Berliner 2002; Au 2007; Hamilton u.a. 2007; Herman 2004; Jacob 2005; Jacob/Levitt 2003; Nichols/Berliner 2007; Ryan u.a. 2007; Schwartz Chrismer/Hodge/Saintil 2006; Stecher 2002). Unter weniger restriktiven und punitiven Bedingungen können sich hingegen eher produktive Handlungs- und Erlebensmuster bei Schüler/innen, Lehrpersonen und Schulleitungen einstellen und funktionale Entwicklungen zur Steigerung der Qualität des Unterrichts beobachtet werden (vgl. Abrams/Madaus 2003; Brozo/Hargis 2003). Auf die Bedeutung des Bildungsmonitoringkontextes für die Bestimmung der Funktionalität von zentralen Prüfungen deuten auch die Analysen von Fuchs/Wößmann (vgl. 2007) hin, in denen die Effekte zentraler Prüfungen in Abhängigkeit anderer Steuerungselemente wie beispielsweise Schulautonomie untersucht worden sind. Hamilton u.a. (vgl. 2007) wiederum fanden in ihren vergleichenden Analysen zur Implementation und zu den Effekten standardbezogener Accountability-Systemen deutliche Unterschiede zwischen den Staaten, die sie in den unterschiedlichen Konzeptionen begründet sehen. Diese sind zudem auch fachspezifisch ausgeprägt (vgl. Hamilton u.a. 2007, S. 130).

Die Implementation neuer output-orientierter Steuerungsstrukturen im Schulwesen kann zurzeit idealtypisch im Rahmen der Einführung zentraler Abiturprüfungen in verschiedenen Bundesländern, in dieser Studie in Bremen und Hessen, untersucht werden.[1] In zweifacher Hinsicht wird es möglich sein, die Effekte der Einführung unter differenzieller Perspektive zu untersuchen. Zum einen hat das Bundesland Bremen zentrale Abiturprüfungen schrittweise für die drei schriftlichen Prüfungsfächer eingeführt, wobei im ersten Jahr der Durchführung (2007) einzig das dritte Prüfungsfach (Grundkurs), in 2008 auch die beiden ersten schriftlichen Prüfungsfächer (Leistungskurse) zentral geprüft werden. Der Vergleich der Effekte in Grund- und Leistungskursen wird sodann in Bremen zu systematischen Erkenntnissen hinsichtlich der Effekte der Implementation innerhalb eines Jahrganges und zwischen den verschiedenen Abiturjahrgängen führen.

Zum anderen können die Effekte in Hessen und Bremen vergleichend miteinander untersucht werden. Im Gegensatz zu Bremen führte Hessen in 2007 gleich zu Beginn in allen drei schriftlichen Prüfungsfächern zentrale Abiturprüfungen ein. Hier sind somit differenzielle Analysen auf Bundeslandebene realisierbar, wobei die auf der gymnasialen Stufe durch die bundesländerübergreifenden Vereinbarungen für die Gestaltung der gymnasialen Oberstufe und der Abiturprüfungen einen solchen empirischen Vergleich trotz länderspezifischer Rahmenbedingungen legitimieren.

1 Zentral vorgegeben sind in diesen beiden Ländern die Abituraufgaben. Die Korrektur und Benotung erfolgt wie im dezentralen Abitur dezentral in den Schulen durch die Lehrer/-innen selber, allerdings mit einem detailliert beschriebenen Erwartungshorizont und Beurteilungsraster.

Im Rahmen dieses Beitrages steht insbesondere die zweite Analysestrategie im Fokus. Die zentrale Fragestellung lautet, inwiefern es auf Schulebene Hinweise zu Effekten der Einführung zentraler Abiturprüfungen auf das verständnisorientierte, unterstützende Lernen der Schüler/innen gibt und inwiefern sich die diesbezügliche Varianz zwischen Schulen im Bundesländervergleich unterscheidet.

1. Theoretisches Analysemodell, aktueller Forschungsstand und Hypothesenbildung

1.1 Educational Governance

Die Analyse der Effekte neuer Steuerungsstrukturen und -mechanismen lässt sich auf dem Hintergrund der Educational Governance-Forschung realisieren. Dabei wird das Bildungssystem als Mehrebenensystem beschrieben, in dem verschiedene Akteure, die in einem spezifischen Abhängigkeitsverhältnis zueinander stehen, die Leistungen kollektiv bzw. ko-produktiv erzeugen (vgl. Altrichter/Brüsemeister/Wissinger 2007; Kussau/Brüsemeister 2007b; Schimank 2007). Dies führt dazu, dass Steuerung nur innerhalb eines doppelten indirekten Verhältnisses zwischen den Akteuren denkbar ist. Steuerung ist damit nicht ein deterministischer top-down Prozess, bei dem mit Sicherheit die Vorgaben und Regelungen von den Akteuren umgesetzt werden, sondern ein probabilistischer top-down und bottom-up-Prozess, wobei Erwartungen nur unter bestimmten Umständen und in Abhängigkeit verschiedener mediierender Faktoren erreicht werden können. Untersucht wird dabei, wie die verschiedenen Akteure im Mehrebenensystem die verschiedenen Abhängigkeiten bearbeiten und wie sich eine veränderte Regelungsstruktur auf die Leistungsstruktur im Mehrebenensystem auswirkt (vgl. Altrichter/Heinrich 2007; Kussau/Brüsemeister 2007a). Ergebnisse aus der Implementationsforschung verweisen auf differente Mechanismen der Verarbeitung externer Vorgaben. So formulieren Spillane/Reiser/Reimer (vgl. 2002), dass in den Schulen nicht nur eine mit den Zielen der Innovation kongruente, sondern auch mit den Zielen weniger kompatible Umsetzung realisiert wird. Fend (vgl. 2006) bezeichnet die entsprechenden Mechanismen als Rekontextualisierung. „Rekontextualisierung meint deshalb Handeln im Rahmen von Ordnungen des Zusammenhandelns angesichts gegebener Umwelten, vermittelt durch die Selbstreferenz, die Interessen und Ressourcen der Handelnden" (Fend 2006, S. 181). Dies bedeutet für die Implementation zentraler Abiturprüfungen, dass diese topdown im Bildungswesen eingeführte administrative Reform aufgrund komplexer Transformationsprozesse je nach Schulkultur, professionellen Orientierungen, individuellen Handlungsabsichten oder Ressourcen der Handelnden unterschiedlich aufgenommen, abgewehrt oder umgeformt werden. Dies impliziert, dass sich diese Transformationsprozesse im Sinne vorstrukturierender Prozessfaktoren auf spezifische Dimensionen schulischen Handelns sowie auf die Lernleistungen der Schüler/innen abbilden und Unterschiede zwischen den Schulen erzeugen werden. Im Rahmen der Analysen in diesem Beitrag interessieren insbesondere die Effekte auf das kognitiv aktivie-

rende, verständnisorientierte und unterstützende Lernen der Schüler/innen. Diese Aspekte haben sich im Kontext der Lehr-Lernforschung als zentrale Qualitätsindikatoren herausgestellt (vgl. Klieme 2006) und wurden in bisherigen Studien zur Analyse der Effekte zentraler Abiturprüfungen erst ansatzweise auf individueller Ebene untersucht, so dass Erkenntnisse darüber fehlen, inwiefern sich Schulen darin unterscheiden, ein für den Lernprozess der Schüler/innen funktionales bzw. weniger funktionales Unterrichtshandeln zu realisieren.

1.2 Effekte zentraler Prüfungen auf das Lernen der Schüler/innen

Auf individueller Ebene verweisen die Ergebnisse der TIMSS-Analysen von Baumert/Watermann (vgl. 2000) darauf, dass sich das Zentralabitur nicht negativ auf das individuelle Lern- und Motivationsgeschehen der Schüler/innen auswirkt und vor allem im unteren Leistungsbereich standardsichernd wirken kann. Allerdings sind die Ergebnisse fachspezifisch zu betrachten: Ein verständnisorientiertes Lernen im Mathematikunterricht wird tendenziell häufiger von Schülerinnen und Schülern berichtet, die unter den Bedingungen eines Zentralabiturs lernen. Im Physikunterricht zeigen sich zwischen den beiden Prüfsystemen keine Unterschiede. Andere Fächer wurden nicht untersucht. Die Studien von Bishop (vgl. 1999 S. 391) auf der Basis internationaler Vergleichsdaten weisen ebenfalls darauf hin, dass das Lernen der Schülerinnen und Schüler in Ländern mit zentralen Abschlussprüfungen nicht mit einem geringeren Ausmaß an kognitiver Aktivierung im Unterricht einhergeht. Vielmehr zeigt sich, dass Schüler/innen in Ländern mit zentralen Abschlussprüfungen weniger häufig Memorieren als notwendige Strategie betrachten, um mathematisches und naturwissenschaftliches Wissen zu erwerben, Lehrpersonen häufiger mit ihren Schüler/innen Experimente im Unterricht durchführen und dass zentrale Abschlussprüfungen zudem die Durchführung von Gruppenarbeiten oder die zeitliche Intensität in der Bearbeitung von mathematischen Problemen nicht einschränken.

Die eigenen Analysen lassen vermuten, dass die Einführung zentraler Abiturprüfungen einzig im dritten Prüfungsfach, wie dies Bremen im ersten Jahr der Durchführung realisiert worden ist, zu einer Stärkung der kognitiven Aktivierung und Unterstützung durch die Lehrpersonen in den Grundkursen geführt hat (vgl. Maag Merki/Holmeier 2008). Dieser Effekt scheint auszubleiben, wenn, wie in Hessen, alle drei schriftlichen Prüfungsfächer zentral geprüft werden. Insbesondere zeigt sich in Hessen auch unter zentralem Prüfsystem die bereits in der TIMS-Studie (vgl. Baumert/Köller 2000a, 2000b) festgestellte substanzielle Differenz zwischen dem Niveau der kognitiven Aktivierung in den Grund- und Leistungskursen – mit einem höheren Anteil an kognitiver Aktivierung in Leistungskursen – während dem diese Differenz in Bremen nicht mehr signifikant ist.

1.3 Fragestellung und Hypothesen

Es stellt sich die Frage, inwiefern sich die auf individueller Ebene festgestellten Effekte auch auf Schulebene abbilden. Unter Berücksichtigung der Ergebnisse und Theorien der Implementationsforschung (vgl. Spillane/Reiser/Reimer 2002; Fend 2006) kann davon ausgegangen werden, dass es zwischen den Schulen innerhalb eines Bundeslandes systematische Differenzen hinsichtlich der Unterrichtsgestaltung in den zentral geprüften Kursen gibt. Die Ergebnisse von Baumert/Watermann (vgl. 2000), die auf standardisierende Effekte zentraler Prüfungen in den Grundkursen hinweisen, lassen zudem vermuten, dass die Unterschiede besonders im dritten Prüfungsfach (Grundkurs) sichtbar werden. Für die Überprüfung dieser Hypothese dienen insbesondere die Analysen in Hessen, wo zentrale Abiturprüfungen sowohl in den Leistungs- wie auch in den Grundkursen eingeführt worden sind.

Da in Bremen einzig das dritte schriftliche Prüfungsfach zentral geprüft wird, in Hessen aber alle drei schriftlichen Prüfungsfächer, besteht aufgrund der Ergebnisse von Hamilton u.a. (vgl. 2007) des Weiteren die Annahme, dass sich diese unterschiedlichen Konzeptionen in unterschiedlichen Ergebnisstrukturen zwischen den Bundesländern abbilden werden, wobei die Varianz in Bremen aufgrund des Vorfindens von zentral und dezentral geprüften Prüfungsfächern größer sein wird als in Hessen.

2. Forschungsdesign und Stichprobe

Basis der nachfolgenden Auswertungen sind schriftliche standardisierte Befragungen von Schülerinnen und Schülern aus vier Kursen pro Gymnasium mit gymnasialer Oberstufe (je ein Grund- und Leistungskurs in Mathematik und Englisch), die kurz vor den schriftlichen Abiturprüfungen im Frühjahr 2007 in den Schulen bzw. in den jeweiligen Kursen durchgeführt worden sind. Im Falle mehrerer entsprechender Kurse pro Schule erfolgte die Auswahl zufällig.

Von den insgesamt 20 Gymnasien mit gymnasialer Oberstufe in Bremen haben sich an den Erhebungen bei den Schüler/innen vor dem Abitur 18 Schulen beteiligt. Bei den Schülerinnen und Schülern konnte ein Rücklauf von 49,6% (N = 751) realisiert werden. In Hessen beteiligten sich 18 Gymnasien mit gymnasialer Oberstufe an den Erhebungen. Diese Schulen wurden nach bestimmten Kriterien ausgesucht (Region, Stadt-Land, Größe der Schule, Profil des Gymnasiums), um eine möglichst repräsentative Stichprobe innerhalb des Bundeslandes zu erhalten. Konkret gehören zur Stichprobe zwölf Gymnasien, zwei Gymnasien nur mit gymnasialer Oberstufe, drei Kooperative Gesamtschulen mit gymnasialer Oberstufe sowie eine Integrierte Gesamtschule mit gymnasialer Oberstufe. Bei den Schülerinnen und Schülern in Hessen beträgt der Rücklauf 67,5% (N = 973). Damit ergibt sich für Bremen und Hessen eine solide Datengrundlage für die Interpretation der Ergebnisse.

2.1 Erhebungsinstrumente

Zur Erfassung eines lernförderlichen und kognitiv anspruchsvollen Unterrichts werden zwei zentrale Aspekte berücksichtigt: die kognitive Aktivierung und die wahrgenommene Unterstützung durch die Lehrperson (vgl. Deci/Ryan 1993; Klieme 2006). Insgesamt wurden dazu vier Skalen eingesetzt, wobei die Schülerinnen und Schüler die einzelnen Items jeweils für ihre drei Kurse eingeschätzt haben, die sie als ihre drei schriftlichen Prüfungsfächer für das Abitur gewählt haben. Damit wird es möglich sein, die Unterrichtsgestaltung in den Leistungskursen (1. und 2. Prüfungsfach) mit der Unterrichtsgestaltung in den Grundkursen (3. Prüfungsfach) zu vergleichen. Differenzielle Analysen haben gezeigt, dass durch das gewählte Erhebungsdesign – Beschränkung auf Grund- und Leistungskursen in den Fächern Englisch und Mathematik – dennoch alle Prüfungsfächer repräsentativ entsprechend der Grundgesamtheit abgebildet werden können. Die einzelnen Items wurden von den Schüler/innen auf einer viergestuften Antwortskala eingeschätzt (1=trifft gar nicht zu … 4=trifft genau zu).

Elaboration (4 Items): „Wir haben im Unterricht immer wieder Gelegenheit, die im Fach erworbenen Kenntnisse mit Kenntnissen aus anderen Fächern zu verknüpfen." Cronbachs Alpha = .73 (Bremen)/.70 (Hessen); Quelle: Leutwyler/Maag Merki (vgl. 2005)

Kompetenzunterstützung (5 Items): „Im Unterricht informiert mich die Lehrperson regelmäßig über meine Fortschritte." Cronbachs Alpha = .80 (Bremen)/.76 (Hessen); Quelle: Prenzel u.a. (vgl. 1996).

Autonomieunterstützung (4 Items): „Im Unterricht habe ich die Möglichkeit, neue Themen selbstständig zu erkunden." Cronbachs Alpha = .68 (Bremen)/.67 (Hessen); Quelle: Prenzel u.a. (vgl. 1996).

Motivierungsfähigkeit/inhaltliches Interesse (5 Items): „Unsere Lehrperson kann Schülerinnen und Schüler manchmal richtig begeistern." Cronbachs Alpha = .81 (Bremen)/.82 (Hessen); Quelle: Leutwyler/Maag Merki (vgl. 2005)

2.2 Auswertungsstrategien

Um Schulunterschiede zu untersuchen, wurden Latent Class Analysen mit Hilfe des Statistikpaketes „Latent GOLD 4.0" (vgl. Vermunt/Magidson 2005) durchgeführt. Latent Class-Analysen sind multivariate probabilistische Klassifikationsverfahren, die es erlauben, Personen/Objekten (in diesen Analysen: „Schulen") in latente Klassen kriterienbasiert zu kategorisieren. Dabei wird davon ausgegangen, dass Klassen disjunkt und exhaustiv sind, jede Person/jedes Objekt einer Klasse angehört und nur einer Klasse angehören kann (vgl. Rost 2004). Für die Identifikation der angemessenen Anzahl von Klassen wurde das Bayes'sche Informationskriterium (BIC) herangezogen (vgl. Rost 2004). Niedrigere Werte beim BIC weisen auf eine bessere Modellanpassung hin. Zusätzlich wird der Klassifikationsfehler als Gütekriterium herangezogen, der es erlaubt, die Wahrscheinlichkeit einer Fehlzuordnung der Objekte zu den einzelnen Clustern zu bestim-

men. Basis für die Analysen sind die auf Schulebene aggregierten Individualdaten der Schüler/innen zu den in Abschnitt 3.1 beschriebenen Indikatoren. Die Daten entsprechen einem kontinuierlichen Skalentyp. Zur detaillierten Beschreibung der Kategorien werden varianzanalytische wie auch inferenzstatistische Verfahren herangezogen.

3. Ergebnisse

3.1 Varianz zwischen den Schulen

In beiden Bundesländern unterscheiden sich die Schulen in allen untersuchten Dimensionen signifikant voneinander (vgl. Tabelle 1). Der Vergleich zwischen Grund- und Leistungskursen innerhalb eines Bundeslandes zeigt, dass der Anteil an erklärter Varianz aufgrund der Schulzugehörigkeit insbesondere beim Indikator „Elaboration" in den Grundkursen höher liegt als in den anderen Dimensionen. In Bremen beträgt der Varianzanteil in dieser Dimension aufgrund der Schulzugehörigkeit 8.1%, in Hessen 7.3%.

Tab. 1: **Anteil an erklärter Varianz zwischen den Schulen; Einfaktorielle univariate Varianzanalyse**

	Hessen		**Bremen**	
	Leistungskurs	Grundkurs	Leistungskurs	Grundkurs
Elaboration	F = 3,47*** (3,0%)	F = 4,34*** (7,3%)	F = 5,47*** (5,9%)	F = 3,79*** (8,1%)
Motivierungsfähigkeit	F = 4,80*** (4,1%)	F = 3,51*** (5,9%)	F =6,34*** (6,8%)	F = 2,58*** (5,7%)
Autonomieunterstützung	F = 6,31*** (5,3%)	F = 3,04*** (5,2%)	F = 5,60*** (6,0%)	F = 2,92*** (6,4%)
Kompetenzunterstützung	F = 6,24*** (5,3%)	F = 3,29*** (5,6%)	F = 3,83*** (4,2%)	F = 1,87* (4,2%)
*** p < .000, ** p < .01, * p < .05				

3.2 Latent Class Analysen in Bremen und Hessen

3.2.1 Hessen

In Tabelle 2 sind die Werte zur Modellgüte der Latent Class Analysen in Hessen angegeben. Dabei zeigt sich, dass die Zwei-Klassenlösung die beste Modellgüteanpassung (BIC = –100.35) und gleichzeitig einen geringen Klassifikationsfehler von 0.04% aufweist (vgl. Tabelle 2). Die Wald-Statistiken verweisen außer beim Indikator „Motivierungsfähigkeit/inhaltliches Interesse der Lehrpersonen" im Grundkurs auf signifikante Unterschiede zwischen den beiden Clustern, wobei Cluster 2 jeweils signifikant höhere Werte erreicht als Cluster 1 (vgl. Tabelle 3 bzw. Abbildung 1).

Tab. 2: **Vergleich der Modellgüteanpassung der verschiedenen Lösungen in Hessen**

		LL	BIC(LL)	Npar	Class. Err.
Model1	1-Cluster	68.6819	-91.1178	16	0
Model2	2-Cluster	97.8684	-100.3545	33	0.0004
Model3	3-Cluster	115.4406	-86.3626	50	0.0141
Model4	4-Cluster	133.9046	-74.1543	67	0.0006
Model5	5-Cluster	152.1741	-61.557	84	0.0037
Model6	6-Cluster	165.9702	-40.0129	101	0.0013
Model7	7-Cluster	180.5406	-20.0173	118	0.0001

LL = Log-Likelihood; BIC (LL): Bayes'sche Informationskriterium;
Npar = number of parameters; Class. Err. = Classification Error

Tab. 3: **Kennwerte der beiden Cluster in Hessen**

1. und 2. Prüfungsfach (Leistungskurse)	**Cluster**	**N**	**M**	**SD**	**Wald**	**E**
Elaboration	1. geringe kogn. Aktivierung GK	1354	2.49	0.70	11.06 ***	0.20
	2. mittlere kogn. Aktivierung GK	560	2.63	0.68		
Motivierungsfähigkeit/ inhaltliches Interesse	1. geringe kogn. Aktivierung GK	1355	2.65	0.71	15.54 **	0.30
	2. mittlere kogn. Aktivierung GK	570	2.86	0.67		
Autonomieunterstützung	1. geringe kogn. Aktivierung GK	1352	2.51	0.63	4.31 *	0.25
	2. mittlere kogn. Aktivierung GK	569	2.66	0.62		
Kompetenzunterstützung	1. geringe kogn. Aktivierung GK	1353	2.55	0.63	5.97 *	0.26
	2. mittlere kogn. Aktivierung GK	568	2.71	0.68		
3. Prüfungsfach (Grundkurse)	**Cluster**	**N**	**M**	**SD**	**Wald**	**E**
Elaboration	1. geringe kogn. Aktivierung GK	678	2.17	0.74	45.48 **	0.50
	2. mittlere kogn. Aktivierung GK	281	2.53	0.70		
Motivierungsfähigkeit/ inhaltliches Interesse	1. geringe kogn. Aktivierung GK	677	2.60	0.67	2.95 n.s.	0.15
	2. mittlere kogn. Aktivierung GK	285	2.70	0.66		
Autonomieunterstützung	1. geringe kogn. Aktivierung GK	677	2.37	0.63	4.07 *	0.18
	2. mittlere kogn. Aktivierung GK	285	2.48	0.59		
Kompetenzunterstützung	1. geringe kogn. Aktivierung GK	676	2.53	0.63	6.33 *	0.26
	2. mittlere kogn. Aktivierung GK	284	2.69	0.59		

*** $p < .000$, ** $p < .01$, * $p < .05$

M = Mittelwert, SD = Standardabweichung, Wald = Testwert zur Prüfung der Unterschiede zwischen den Clustern auf Signifikanz, E = Effektgrößen (Cohen, 1988)

Die Effektgrößen (vgl. Cohen 1988) weisen mit Werten zwischen d = 0.15 und d = 0.30 in den meisten Dimensionen auf schwache Effekte hin. Einzig der Unterschied in der kognitiven Aktivierung im Grundkurs („Elaboration_GK") ist mit d = 0.50 substanziell. Aus diesem Grund werden die beiden Cluster als „geringe kognitive Aktivierung in Grundkursen" (Cluster 1) und „mittlere kognitive Aktivierung in Grundkursen" (Cluster 2) charakterisiert. Zu Cluster 1 können zwölf Schulen zugeordnet werden, zu Cluster 2 sechs Schulen.

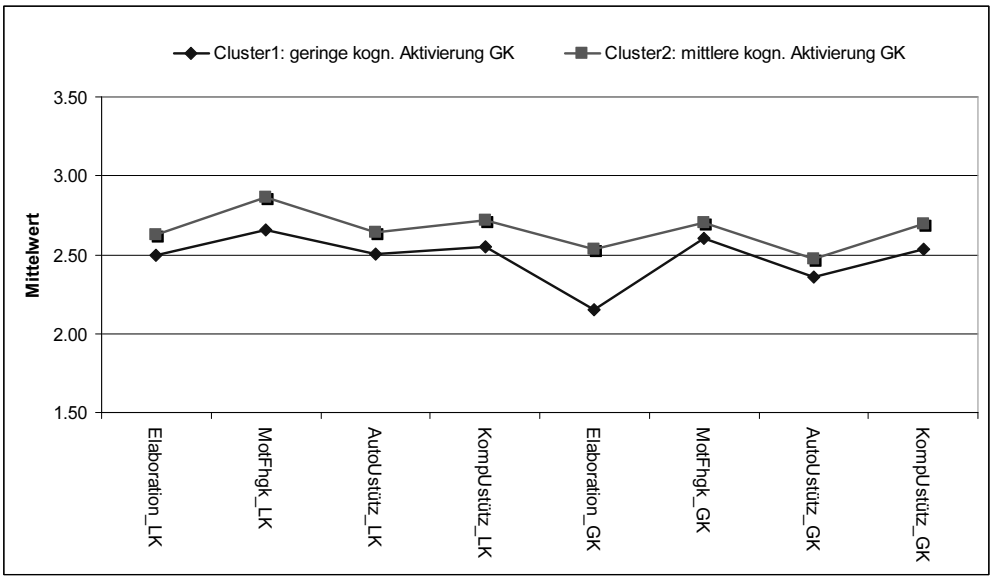

Abb. 1: Ergebnis der LCA in Hessen; LK=Leistungskurs; GK=Grundkurs; Unterrichtsdimensionen: Elaboration, Motivierungsfähigkeit / inhaltliches Interesse (MotFhgk), Autonomieunterstützung (AutoUstütz), Kompetenzunterstützung (KompUstütz)

Innerhalb der beiden Cluster zeigen sich auf der Basis von T-Tests für unabhängige Stichproben systematische Unterschiede zwischen den erfassten Unterrichtsqualitäten in den Leistungs- und Grundkursen. Die Werte für die Leistungskurse in Cluster 1 („geringe kognitive Aktivierung in Grundkursen") sind insbesondere beim Indikatoren „Elaboration" hochsignifikant höher als in den Grundkursen (d = 0.46, $p < .000$). Schüler/innen in Schulen dieses Clusters erleben damit in den Leistungskursen eine systematisch stärkere kognitive Aktivierung (M = 2.49, SD = 0.70) als in den Grundkursen (M = 2.17, SD = 0.74). Ebenfalls nehmen sie in den Leistungskursen bei ihren Lehrpersonen eine deutlich höhere Autonomieunterstützung wahr (M = 2.51, SD = 0.63) als in den Grundkursen (M = 2.37, SD = 0.63) (d = 0.22, $p < .000$). Hinsichtlich der erlebten Motivierungsfähigkeit und Kompetenzunterstützung zeigen sich keine signifikanten Unterschiede zwischen den Grund- und Leistungskursen (vgl. Tabelle 3). In Cluster 2 („Mittlere kognitive Aktivierung in Grundkursen") ist der Unterschied zwischen

Grund- und Leistungskursen hinsichtlich der kognitiven Aktivierung zwar nach wie vor signifikant, allerdings schwach ($d = 0.15$, $p < .05$). Etwas stärker sind die Unterschiede bei den Indikatoren „Motivierungsfähigkeit" ($d = 0.24$, $p < .000$) und „Autonomieunterstützung" ($d = 0.31$, $p < .000$). Schüler/innen in Schulen, die dem Cluster 2 angehören, erleben damit im Durchschnitt in den Leistungskursen eine stärker ausgeprägte Motivierungsfähigkeit sowie Unterstützungsqualität als in den Grundkursen.

3.2.2 Bremen

Die Latent Class Analysen für Bremen zeigen für die 3-Klassen-Lösung die beste Modellanpassung (BIC = -29.72) (vgl. Tabelle 4). Die Zuteilung der Schulen zu den drei Clustern erfolgt mit einem sehr geringen Klassifikationsfehler von 0.06%.

Tab. 4: **Vergleich der Modellgüteanpassung der verschiedenen Lösungen in Bremen**

		LL	BIC(LL)	Npar	Class. Err.
Model1	1-Cluster	27.1921	-8.1382	16	0.0000
Model2	2-Cluster	58.6174	-21.8526	33	0.0019
Model3	3-Cluster	87.1216	-29.7247	50	0.0006
Model4	4-Cluster	100.4985	-7.3420	67	0.0004
Model5	5-Cluster	121.3583	0.0745	84	0.0023
Model6	6-Cluster	133.2731	25.3814	101	0.0003
Model7	7-Cluster	142.5741	55.9157	118	0.0013

LL = Log-Likelihood; BIC (LL): Bayes'sche Informationskriterium;
Npar = number of parameters; Class. Err. = Classification Error

Die Kennwerte der drei Cluster sind in Tabelle 5 zusammengefasst und erlauben eine entsprechende Charakterisierung (vgl. auch Abbildung 2). Schulen in Cluster 1 fallen durch ihre relativ hohen Durchschnittswerte in den Grundkursen im Vergleich zu den Leistungskursen auf. So erreicht Cluster 1 in den beiden Dimensionen „Motivierungsfähigkeit/inhaltliches Interesse" und „Kompetenzunterstützung" signifikant höhere Werte in den Grundkursen als in den Leistungskursen (Effektgrößen $d = 0.12$, $p < .05$ bzw. $d = 0.21$, $p < .000$). In den beiden anderen Dimensionen zeigen sich keine signifikanten Unterschiede zwischen den beiden Kurstypen. In diesem Cluster scheint der zentral geprüfte Grundkurs daher im Vergleich zu den dezentral geprüften Leistungskursen stärker fokussiert zu werden. Er wird somit als „Fokussierung auf zentral geprüfte Grundkurse" bezeichnet. Insgesamt können zehn Schulen diesem Cluster zugeordnet werden. Schulen in Cluster 2 weisen im Durchschnitt eher hohe Werte in den Leistungskursen und relativ geringe Werte in den Grundkursen auf, so dass in diesen Schulen der Fokus eher auf den Leistungskursen liegt. Die Unterschiede zwischen Leistungs- und Grundkursen in den untersuchten Dimensionen sind sodann bis auf die Dimension „Kompetenzunterstützung" hoch signifikant und inhaltlich substanziell (Effekt-

größen zwischen d = 0.27 (Motivierungsfähigkeit/inhaltliches Interesse) und d = 0.63 (Elaboration). Damit entspricht das Ergebnis der Schulen in Cluster 2 dem Ergebnisbild, wie es in der TIMS-Studie (vgl. Baumert/Köller 2000a, 2000b) beobachtet worden ist. Dieser Cluster wird daher als „Traditionelles Profil" bezeichnet. Insgesamt können vier Schulen diesem Cluster zugeordnet werden.

Tab. 5: **Kennwerte der drei Cluster in Bremen**

1. und 2. Prüfungsfach (Leistungskurse)	**Cluster**	**N**	**M**	**SD**	**Wald**
Elaboration	1. Fokussierung auf zentral geprüfte GK	981	2.40	0.75	11.41 ***
	2. Traditionelles Profil	310	2.58	0.72	
	3. Fokussierung auf GK und LK	208	2.74	0.73	
Motivierungsfähigkeit / inhaltliches Interesse	1. Fokussierung auf zentral geprüfte GK	981	2.49	0.73	6.82 *
	2. Traditionelles Profil	310	2.59	0.70	
	3. Fokussierung auf GK und LK	208	2.90	0.72	
Autonomieunterstützung	1. Fokussierung auf zentral geprüfte GK	980	2.45	0.68	86.59 ***
	2. Traditionelles Profil	310	2.44	0.61	
	3. Fokussierung auf GK und LK	208	2.82	0.62	
Kompetenzunterstützung	1. Fokussierung auf zentral geprüfte GK	976	2.42	0.72	47.70 ***
	2. Traditionelles Profil	310	2.49	0.63	
	3. Fokussierung auf GK und LK	207	2.74	0.65	
3. Prüfungsfach (Grundkurse)	**Cluster**	**N**	**M**	**SD**	**Wald**
Elaboration	1. Fokussierung auf zentral geprüfte GK	492	2.48	0.73	30.06 ***
	2. Traditionelles Profil	155	2.13	0.71	
	3. Fokussierung auf GK und LK	103	2.58	0.74	
Motivierungsfähigkeit/ inhaltliches Interesse	1. Fokussierung auf zentral geprüfte GK	491	2.57	0.66	14.51 ***
	2. Traditionelles Profil	155	2.40	0.69	
	3. Fokussierung auf GK und LK	102	2.79	0.61	
Autonomieunterstützung	1. Fokussierung auf zentral geprüfte GK	490	2.45	0.62	34.89 ***
	2. Traditionelles Profil	155	2.23	0.63	
	3. Fokussierung auf GK und LK	102	2.67	0.57	
Kompetenzunterstützung	1. Fokussierung auf zentral geprüfte GK	489	2.57	0.65	14.47 ***
	2. Traditionelles Profil	155	2.42	0.69	
	3. Fokussierung auf GK und LK	102	2.78	0.65	

*** p < .000, ** p < .01, * p < .05

M = Mittelwert, SD = Standardabweichung, Wald = Testwert zur Prüfung der Unterschiede zwischen den Clustern auf Signifikanz, GK = Grundkurs, LK = Leistungskurs

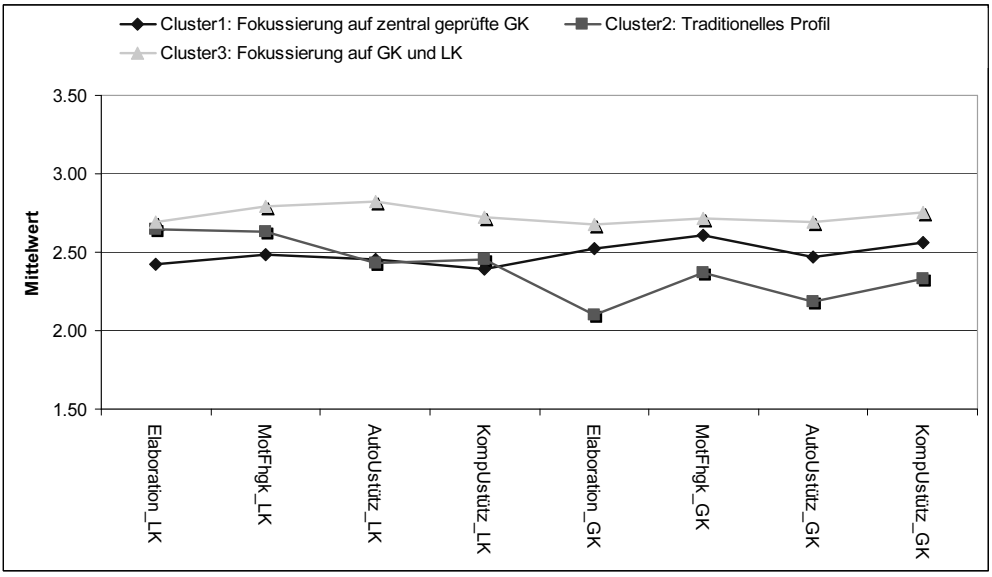

Abb. 2: Ergebnis der LCA in Bremen; LK=Leistungskurs; GK=Grundkurs; Unterrichtsdimensionen: Elaboration, Motivierungsfähigkeit / inhaltliches Interesse (MotFhgk), Autonomieunterstützung (AutoUstütz), Kompetenzunterstützung (KompUstütz)

Schulen in Cluster 3 unterscheiden sich von den beiden anderen Clustern insofern, als diese im Durchschnitt sowohl in den Leistungs- als auch in den Grundkursen hohe Werte in allen Dimensionen aufweisen. Die Unterschiede zwischen Grund- und Leistungskursen sind bis auf eine Dimension nicht signifikant: Schüler/innen in Schulen des Clusters 3 erleben in den Leistungskursen eine signifikant größere Autonomieunterstützung als in den Grundkursen ($d = 0.26$, $p < .05$). Aufgrund der relativ gleichen Gewichtung von Grund- und Leistungskursen wird dieser Cluster als „Fokussierung auf Grund- und Leistungskurse" bezeichnet. Es können vier Schulen diesem Cluster zugeordnet werden.

Die Wald-Statistiken zeigen, dass in allen untersuchten Dimensionen signifikante Unterschiede zwischen den drei Clustern bestehen (vgl. Tabelle 5), wobei in allen Dimensionen Cluster 3 die höchsten Mittelwerte erreicht. Zudem zeigen sich die größeren Unterschiede zwischen den Clustern in den Dimensionen zur Erfassung der Unterrichtsqualität in den zentral geprüften Grundkursen.

In den nicht zentral geprüften Leistungskursen führt vor allem die Fokussierung auf Leistungs- *und* Grundkurse (Cluster 3) zu besseren Werten. Zwischen Cluster 1 (Fokussierung auf zentral geprüfte Grundkurse) und Cluster 3 zeigen sich substanzielle Effekte in allen Dimensionen: „Elaboration" ($d = 0.46$, $p < .000$), „Motivierungsfähigkeit/inhaltliches Interesse" ($d = 0.57$, $p < .000$), „Autonomieunterstützung" ($d = 0.55$, $p < .000$) und „Kompetenzunterstützung" ($d = 0.45$, $p < .000$). Cluster 2 (Traditionelles Profil) und Cluster 3 unterscheiden sich mit einer Effektgröße zwischen $d = 0.23$ („Ela-

boration“) und d = 0.62 („Autonomieunterstützung“) ebenfalls in allen Dimensionen signifikant. Cluster 1 und 2 unterscheiden sich einzig in der Dimension „Elaboration“ signifikant (d = 0.24, p < .01).

Dieses Bild zeigt sich auch weitgehend in den Grundkursen. Cluster 3 (Fokussierung auf Grund- und Leistungskurse) erreicht in allen Dimensionen die höchsten Mittelwerte, allerdings in der Dimension „Elaboration“ zusammen mit Cluster 1 (Fokussierung auf zentral geprüfte Grundkurse). Die Mittelwertsdifferenz zwischen Cluster 2 (Traditionelles Profil) und den Clustern 1 bzw. 3 ist mit einer Effektgröße von d= 0.48 (p < .000) bzw. d = 0.62 (p < .000) zudem inhaltlich substanziell. In den anderen Dimensionen sind die Differenzen zwischen Cluster 1 und 3 ebenfalls signifikant und variieren zwischen d = 0.33 („Motivierungsfähigkeit/inhaltliches Interesse“) und d = 0.35 („Autonomieunterstützung“).

Die Effekte zwischen Cluster 2 (Traditionelles Profil) und Cluster 3 (Fokussierung auf Grund- und Leistungskurse) sind ebenfalls signifikant, sind aber stärker: „Motivierungsfähigkeit/inhaltliches Interesse“ d = 0.58 (p < .000), „Autonomieunterstützung“ d = 0.72 (p < .000), „Kompetenzunterstützung“ d = 0.55 (p < .000).

Zwischen Cluster 1 (Fokussierung auf zentral geprüfte Grundkurse) und Cluster 2 (Traditionelles Profil) sind die Unterschiede ebenfalls signifikant, allerdings geringer: „Motivierungsfähigkeit/inhaltliches Interesse“ d = 0.25 (p < .000), „Autonomieunterstützung“ d = 0.37 (p < .000), „Kompetenzunterstützung“ d = 0.23 (p < .000).

In den zentral geprüften Grundkursen kann somit durch eine verstärkte Fokussierung auf die Grundkurse *mit oder ohne* gleichzeitiger Fokussierung auf die Leistungskurse eine in Teilen höhere Unterrichtsqualität erreicht werden, als wenn einzig auf die Leistungskurse fokussiert wird (Cluster 2). Dies ist insbesondere der Fall hinsichtlich der Realisierung eines kognitiv anspruchsvollen Unterrichts („Elaboration“).

4. Diskussion

In diesem Beitrag wurde untersucht, inwiefern es auf Schulebene Hinweise zu Effekten der Einführung zentraler Abiturprüfungen auf das verständnisorientierte, unterstützende Lernen der Schüler/innen gibt und inwiefern sich die diesbezügliche Varianz zwischen Schulen im Bundesländervergleich unterscheidet.

In einem ersten Schritt wurden die Unterschiede zwischen den Schulen innerhalb der beiden Bundesländer untersucht. Dabei wurde die Hypothese aufgestellt, dass es zwischen den Schulen innerhalb eines Bundeslandes systematische Differenzen hinsichtlich der Unterrichtsgestaltung in den zentral geprüften abiturrelevanten Kursen geben wird, wobei die Unterschiede insbesondere im dritten Prüfungsfach (Grundkurs) sichtbar werden.

Die Ergebnisse von Hessen, wo alle drei schriftlichen Prüfungsfächer zentral geprüft worden sind, bestätigen in der Tendenz diese Erwartungen. Sie zeigen, dass sich die Schulen hinsichtlich der wahrgenommenen Unterrichtsqualität in den abiturrelevanten Prüfungsfächern systematisch unterscheiden, wobei aufgrund der durchgeführten Va-

rianzanalysen die größten Unterschiede zwischen den Schulen im Ausmaß der kognitiven Aktivierung im Grundkurs bestehen.

In Hessen lassen sich zwei Umsetzungstypen unterscheiden. Der größte Unterschied zwischen den beiden Clustern wird entsprechend den Erwartungen (vgl. Baumert/Watermann 2000) im Grundkurs sichtbar, dies allerdings einzig in der kognitiven Aktivierung, nicht aber in den anderen Unterrichtsdimensionen. In beiden Umsetzungstypen zeigt sich analog zu den TIMSS-Analysen (vgl. Baumert/Köller 2000a, 2000b) ein traditionelles Ergebnisbild in dem Sinne, als die Schüler/innen in Leistungskursen einen kognitiv anspruchsvolleren Unterricht sowie eine größere Unterstützungsqualität erleben als in den Grundkursen. Während dem diese Differenzen in Cluster 1, dem zwei Drittel der Schulen zugeordnet worden sind, mit mindestens einer halben Standardabweichung zwischen Leistungs- und Grundkursen substanziell sind, sind die Effekte in Cluster 2 (ein Drittel der Schulen) bedeutsam geringer und in der Dimension „Elaboration" klein.

Die Umsetzung der zentralen Vorgaben scheint somit innerhalb von Hessen mit unterschiedlichen Typen der Unterrichtsgestaltung auf Schulebene assoziiert zu sein. Inwiefern dieses Ergebnis tatsächlich auf einen Effekt der Einführung zentraler Abiturprüfungen zurückzuführen ist, kann an dieser Stelle nicht abschließend beantwortet werden. So könnte es auch sein, dass die Schulen in Cluster 2 bereits vor Einführung zentraler Abiturprüfungen in einem stärkeren Ausmaß einen kognitiv anspruchvollen Unterricht im Grundkurs realisiert haben. Wenngleich sich diese Frage für Hessen erst mit den nächsten Erhebungen in 2008 und 2009 beantworten lassen wird, zeigt der Vergleich mit Bremen zum einen eine interessante Parallele, zum anderen einen bedeutsamen Kontrast, die für die Fragestellung aufschlussreich sind.

So lassen sich in den dezentral geprüften Leistungskursen in Bremen ebenfalls nur zwei Umsetzungstypen, analog zu den Ergebnissen in Hessen, identifizieren: Cluster 1 (Fokussierung auf zentral geprüfte Grundkurse) und Cluster 2 (Traditionelles Profil) unterscheiden sich in den untersuchten Indikatoren in den Leistungskursen nicht, während dem sie sich aber systematisch von Cluster 3 unterscheiden (Fokussierung auf Grund- *und* Leistungskurse). In den zentral geprüften Grundkursen hingegen können in Bremen drei verschiedene Typen der Unterrichtsgestaltung beobachtet werden, während dem dies in Hessen einzig zwei Typen sind. So entspricht Cluster 2 in Bremen einem eher traditionellen Ergebnisprofil, in dem – vergleichbar mit den Ergebnissen der TIMS-Studie (vgl. Baumert/Köller 2000a, 2000b) und analog zu den Ergebnissen in Hessen – im Durchschnitt die Unterrichtsqualität von den Schüler/innen in den Leistungskursen ausgeprägter wahrgenommen wird als in den Grundkursen. Bei den anderen beiden Cluster in Bremen hingegen zeigen sich systematisch differente Muster: In Cluster 3, dem vier Schulen zugeordnet werden konnten, ergeben sich auf hohem Niveau praktisch keine Unterschiede zwischen den Leistungs- und Grundkursen, in Cluster 1, dem zehn Schulen zugeteilt worden sind, zeigt sich sogar in einzelnen Dimensionen eine ausgeprägtere Unterrichtsqualität in den Grundkursen, allerdings mit einer etwas geringer ausgeprägten Unterrichtsqualität in den Leistungskursen. Damit wird in diesen Schulen in stärkstem Maße ein möglicher Teaching-to-the-test-Effekt sichtbar

(vgl. Au 2007; Hamilton u.a. 2007; Jacob 2005; Jacob/Levitt 2003; Nichols/Berliner, 2007), der für die Grundkurse positiv, für die Leistungskurse hingegen weniger positiv zu beurteilen ist. Das für das Lernen produktivste Ergebnisbild, eine gleichgewichtige Fokussierung auf Grund- und Leistungskurse auf relativ hohem Niveau, kann einzig in den Schulen in Cluster 3 beobachtet werden.

Interessant ist der Vergleich der Cluster 1 und 2 in Bremen, die sich in den dezentral geprüften Leistungskursen nicht voneinander unterscheiden, während dem sie dies in den zentral geprüften Grundkursen substanziell tun. Die Einführung zentraler Vorgaben, die innerhalb eines Bundeslandes für alle gleich sind, scheint somit in den Schulen entsprechend den Erwartungen mit unterschiedlichen Umsetzungsformen verknüpft zu sein. Welche Faktoren oder Rekontextualisierungsmechanismen diese unterschiedlichen Realisierungsformen bedingen, kann in weiteren Analysen untersucht werden.

Insgesamt bestätigen die Vergleiche zwischen Bremen und Hessen die aufgestellte Hypothese, dass sich die unterschiedlichen Implementationsmodi in Hessen und Bremen in unterschiedlichen Ergebnisstrukturen zwischen den Bundesländern abbilden, wobei die Varianz in Bremen größer ist als in Hessen.

Die Frage stellt sich, warum sich in den zentral geprüften Grundkursen in Bremen eine größere Varianz zeigt als in den zentral geprüften Grundkursen in Hessen. Eine mögliche Erklärung könnte analog zu den internationalen Erfahrungen (z.B. Hamilton u.a. 2007, S. 130) darin liegen, dass in Bremen aufgrund der Einführung zentraler Abiturprüfungen einzig in den Grundkursen diese im Vergleich zu den Leistungskursen eine besondere Aufmerksamkeit erfahren haben. In Hessen hingegen, wo alle drei schriftlichen Prüfungsfächer zentral geprüft worden sind und somit das Verhältnis zwischen den Grund- und Leistungskursen nicht verändert worden ist, ergibt sich keine neue Gewichtung der beiden Kurstypen. Für Bremen wird sich mit den Ergebnissen der nächsten beiden Jahre zeigen, ob diese besondere Stärkung der Unterrichtsqualität in den Grundkursen auch aufrechterhalten bleiben wird, wenn im Abitur 2008 und 2009 auch die Leistungskurse zentral geprüft werden. Unter der Einschränkung, dass möglicherweise unterschiedliche Rahmenbedingungen in den beiden Bundesländern die differenten Ergebnisstrukturen mitbedingen, bleibt vorerst die These, dass die Einführung zentraler Abiturprüfungen einzig in den Grundkursen zu einer bedeutsamen Stärkung der Unterrichtsqualität in den Grundkursen führt, wobei analog zu den Schulen in Cluster 3 diese Stärkung nicht zwingend mit einer geringeren Gewichtung der Leistungskurse einher gehen muss. Zu untersuchen ist, ob sich dabei fachspezifische Unterschiede ergeben, wie sie aufgrund der Ergebnisse von Hamilton u.a. (vgl. 2007, S. 130) oder Baumert/Watermann (vgl. 2000) erwartbar sind, und inwiefern sich nicht nur die Relation zwischen Leistungs- und Grundkursen, sondern auch das absolute Niveau der Unterrichtsqualität in allen Prüfungsfächern erhöht. Weitergehende Analysen haben zudem zu untersuchen, welche Gründe für diese unterschiedlichen Realisierungsformen vorliegen, damit für die Lehrpersonen und Schulen systematische Weiterbildungs- und Unterstützungsformen aufbereitet werden können, damit der Umgang mit zentralen Vorgaben zu Umsetzungsformen führen, die für das Lernen der Schüler/innen produktiv sind.

Literatur

Abrams, L.M./Madaus, G.F. (2003): The lessons of High-Stakes Testing. In: Educational Leadership 61, H. 3, S. 31–35.

Altrichter, H./Brüsemeister, T./Wissinger, J. (Hrsg.) (2007): Educational Governance. Handlungskoordination und Steuerung im Bildungswesen. Wiesbaden: VS Verlag für Sozialwissenschaften.

Altrichter, H./Heinrich, M. (2007): Kategorien der Governance-Analyse und Transformationen der Systemsteuerung in Österreich. In: Altrichter, H./Brüsemeister T./Wissinger, J. (Hrsg.): Educational Governance. Handlungskoordinationen und Steuerung im Bildungssystem. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 55–98.

Amrein, A.L./Berliner, D.C. (2002): High-stakes testing, uncertainty, and student learning. Education Policy Analysis Archives, 10(18). Verfügbar unter: http://epaa.asu.edu/epaa/v10n18/ (zuletzt eingesehen: 25.9.08).

Au, W. (2007): High-stakes Testing and Curricular Control: A Qualitative Metasynthesis. In: Educational Researcher 36, H. 5, S. 258–267.

Baumert, J./Cordula, A./Klieme, E./Neubrand, M./Prenzel, M./Schiefele, U./ Schneider, W./Tillmann, K.-J./Weiss, M. (Hrsg.) (2003): *PISA 2000* – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland. Opladen: Leske + Budrich.

Baumert, J./Köller, O. (2000a): Motivation, Fachwahlen, selbstreguliertes Lernen und Fachleistungen im Mathematik- und Physikunterricht der gymnasialen Oberstufe. In: Baumert, J./Bos W./Lehmann, B. (Hrsg.): TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe. Opladen: Leske+Budrich, S. 181–213.

Baumert, J./Köller, O. (2000b): Unterrichtsgestaltung, verständnisvolles Lernen und multiple Ziererreichung im Mathematik- und Physikunterricht der gymnasialen Oberstufe. In: Baumert, J./Bos W./Lehmann, B. (Hrsg.): TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe. Opladen: Leske+Budrich, S. 271–315.

Baumert, J./Watermann, R. (2000): Institutionelle und regionale Variabilität und die Sicherung gemeinsamer Standards in der gymnasialen Oberstufe. In: Baumert, J./Bos W./Lehmann, B. (Hrsg.): TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 2. Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe. Opladen: Leske+Budrich, S. 317–372.

Bishop, J.H. (1999): Are national exit examinations important for educational efficiency. In: Swedish Economic Policy Review 6, S. 349–398.

Brozo, W.G./Hargis, C. (2003): Using Low-Stakes Reading Assessment. In: Educational Leadership 61, H. 3, S. 60–64.

Cohen, J. (1988): Statistical power analysis for the behavioral sciences. Hillsdale/NY: Erlbaum.

Deci, E.L./Ryan, R.M. (1993): Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. In: Zeitschrift für Pädagogik 39, H. 2, S. 223–238.

Fend, H. (2006): Neue Theorie der Schule. Einführung in das Verstehen von Bildungssystemen. Lehrbuch. Wiesbaden: Verlag für Sozialwissenschaften.

Fuchs, T./Wößmann, L. (2007): What accounts for international differences in student performance? A re-examination using PISA data. In: Empirical Economics 32, H. 2–3, S. 433–464.

Hamilton, L.S./Stecher, B.M./Marsh, J.A./McCombs, J.S./Robyn, A./Russell, J.L./Naftel, S./Barney, H. (2007): Standards-Based Accountability Under No Child Left Behind. Experiences of Teachers and Administrators in Three States. Santa Monica: Rand.

Herman, J.L. (2004): The Effects of Testing on Instruction. In: Fuhrman, S.H./Elmore, R.F. (Hrsg.): Redesigning Accountability Systems for Education. New York/London: Teachers College Press, S. 141–166.

Jacob, B.A. (2005): Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago public schools. In: Journal of Public Economics 89, H. 5, S. 761–796.

Jacob, B.A./Levitt, S.D. (2003): Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. In: Quarterly Journal of Economics 118, H. 3, S. 843–877.

Klieme, E. (2004): Begründung, Implementation und Wirkung von Bildungsstandards: Aktuelle Diskussionslinien und empirische Befunde. In: Zeitschrift für Pädagogik 50, H. 5, S. 625–634.

Klieme, E. (2006): Empirische Unterrichtsforschung: Aktuelle Entwicklungen, theoretische Grundlagen und fachspezifische Befunde. Einführung in den Thementeil. In: Zeitschrift für Pädagogik 52, H. 6, S. 765–773.

Köller, O./Baumert, J./Cortina, K. S./Trautwein, U./Watermann, R. (2004): Öffnung von Bildungswegen in der Sekundarstufe II und die Wahrung von Standards. In: Zeitschrift für Pädagogik 50, H. 5, S. 679–700.

Köller, O./Baumert, J./Schnabel, K. (1999): Wege zur Hochschulreife: Offenheit des Systems und Sicherung vergleichender Standards. In: Zeitschrift für Erziehungswissenschaft 2, H. 3, S. 385–422.

Kussau, J./Brüsemeister, T. (2007a): Educational Governance: Zur Analyse der Handlungskoordination im Mehrebenensystem der Schule. In: Altrichter, H./Brüsemeister T./Wissinger, J. (Hrsg.): Educational Governance. Handlungskoordination und Steuerung im Bildungswesen. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 15–54.

Kussau, J./Brüsemeister, T. (Hrsg.) (2007b): Governance, Schule und Politik. Zwischen Antagonismus und Kooperation. Wiesbaden: Verlag für Sozialwissenschaften.

Leutwyler, B./Maag Merki, K. (2005): Mittelschulerhebung 2004. Indikatoren zu Kontextmerkmalen gymnasialer Bildung. Perspektiven der Schülerinnen und Schüler: Schul- und Unterrichtserfahrungen. Skalen- und Itemdokumentation. Zürich: Forschungsbereich Schulqualität & Schulentwicklung. Pädagogisches Institut, Universität Zürich.

Maag Merki, K./Holmeier, M. (2008): Die Implementation zentraler Abiturprüfungen. Erste Ergebnisse zu den Effekten der Einführung auf das schulische Handeln der Lehrpersonen. In: E.-M. Lankes (Hrsg.): Pädagogische Professionalität als Gegenstand empirischer Forschung. Münster: Waxmann, S. 233–243.

Nichols, S.L./Berliner, D.C. (2007): Collateral Damage. How Hig-Stakes Testing courrupts American's schools. Cambridge: Harvard Education Press.

Rost, D.H. (2004): Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. In: Zeitschrift für Pädagogik 50, H. 5, S. 662–678.

Ryan, K.E./Ryan, A.M./Arbuthnot, K./Samuels, M. (2007): Students' Motivation for Standardized Math Exams. In: Educational Researcher 36, H. 1, S. 5–13.

Prenzel, M./Kristen, A./Dengler, P./Ettle, R./Beer, T. (1996): Selbstbestimmt motiviertes und interessiertes Lernen in der kaufmännischen Erstausbildung. In: Zeitschrift für Berufs- und Wirtschaftspädagogik. Beiheft 13, S. 108–127.

Schimank, U. (2007): Die Governance-Perspektive: Analytisches Potenzial und anstehende konzeptionelle Fragen. In: Altrichter, H./Brüsemeister T./Wissinger, J. (Hrsg.): Educational Governance. Handlungskoordinationen und Steuerung im Bildungssystem. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 231–261.

Schwartz Chrismer, S./Hodge, S.T./Saintil, D. (Hrsg.) (2006): Assessing NCLB. Perspectives and Prescriptions. Harvard Educational Review 76, H. 4. Cambridge: Harvard Graduate School of Ecucation.

Spillane, J.P./Reiser, B.J./Reimer, T. (2002): Policy Implementation and Cognition: Reframing and Refocusing Implementation Research. In: Review of Educational Research 72, H. 3, S. 387–431.

Stecher, B.M. (2002): Consequences of large-scale, high-stakes testing on school and classroom practice. In: Hamilton, L.S./Stecher B.M./Klein, S.P. (Hrsg.): Making Sense of Test-Based Accountability in Education. Santa Monica: Rand, S. 79–100.

Vermunt, J.K./Magidson, J. (2005): Latent GOLD 4.0 User's Guide. Belmont/Massachusetts: Statistical Innovations Inc.

Abstract: *The introduction of central school-leaving exams is an essential element in the maintainance of a certain standard in the newly implemented output-control models. However, within the German-speaking countries, in particular, we still lack empirical studies showing in how far these exams are functional with regard to achieving the goals set. The present study analyzes the impact of the implementation of central school-leaving exams on essential dimensions of the quality of instruction. This analysis is based on interviews with students from two German Laender – Hesse and Bremen – who took their school-leaving exams in 2007. Through latent class analyses differences in the structures of the result are revealed between the secondary schools and the Laender, which can in part be interpreted as productive with regard to the students´ learning processes.*

Anschrift der Autoren:
Prof. Dr. Katharina Maag Merki, Pädagogische Hochschule Freiburg, Institut für Erziehungswissenschaft, Kunzenweg 21, 79117 Freiburg, E-Mail: maagmerki@ph-freiburg.de
Prof. Dr. Eckhard Klieme, Deutsches Institut für Internationale Pädagogische Forschung, Schloßstrasse 29, 60486 Frankfurt/M., E-Mail: klieme@dipf.de
Dipl. Päd. Monika Holmeier, Deutsches Institut für Internationale Pädagogische Forschung, Schloßstrasse 29, 60486 Frankfurt/M., E-Mail: holmeier@dipf.de